BAB II
TINJAUAN PUSTAKA

2.1. Sistem

Pengertian sistem pada berbagai bidang berbeda-beda, tetapi meskipun istilah sistem yang digunakan bervariasi, namun pada prinsipnya setiap sistem selalu terdiri dari empat elemen, yaitu objek, atribut, hubungan internal dan lingkungan. Serta yang paling penting adalah sistem harus mempunyai tujuan yang akan dicapai secara jelas.

Sistem berasal dari bahasa Latin *(sistema)* dan bahasa Yunani *(sustema)*, adalah suatu kesatuan yang terdiri dari komponen atau elemen yang dihubungkan bersama untuk memudahkan aliran informasi, materi atau energi. Sistem juga merupakan kesatuan bagian-bagian yang saling berhubungan yang berada dalam suatu wilayah serta memiliki item-item penggerak.

Menurut Azhar Susanto (2008 : 18), sistem adalah kumpulan atau grup dari bagian atau komponen apapun baik fisik ataupun non fisik yang saling berhubungan satu sama lain dan bekerja sama secara harmonis untuk mencapai satu tujuan tertentu.

Sedangkan menurut Abdul Kadir (2009:54), sistem adalah sekumpulan elemen yang saling terkait atau terpadu yang dimaksudkan untuk mencapai suatu tujuan.

Dari beberapa definisi diatas, dapat diartikan bahwa sistem adalah sekumpulan unsur atau elemen yang saling berkaitan dan saling mempengaruhi dalam melakukan kegiatan bersama untuk mencapai tujuan.

2.2. Bahasa Pemrograman

Bahasa pemrograman *(programming language)* adalah suatu perangkat lunak dan bahasa yang digunakan untuk membuat program-program komputer

atau sering disebut sebagai bahasa komputer. Bahasa pemrograman menggunakan sistem tata bahasa tertentu atau kata - kata unik untuk dijadikan kode yang bisa menjalankan perintah tertentu pada komputer.

Bahasa pemrograman inilah yang membentuk struktur perangkat lunak sebagai inti dari komputer untuk menjalankan perangkat keras. Tanpa perangkat lunak yang dibuat menggunakan bahasa pemrograman, maka perangkat keras tidak akan berjalan dengan baik atau bahkan tidak dapat berjalan sama sekali.

2.2.1. PHP

Menurut Bunafit Nugroho (2004:139) ada beberapa pengertian tentang PHP. Akan tetapi, kurang lebih PHP dapat kita ambil arti sebagai PHP *Hypertext Preprocesor*, Ini merupakan bahasa yang hanya dapat berjalan pada server yang hasilnya dapat ditampilkan pada klien.

Interpreter PHP dalam mengeksekusi kode PHP pada sisi server (disebut *server-side*) berbeda dengan mesin maya Java yang mengeksekusi program pada sisi klien (*client-side*). PHP merupakan bahasa standar yang digunakan dalam dunia web site. PHP adalah bahasa program yang berbentuk *script* yang diletakkan di dalam server *web.*

Jika dilihat dari sejarah, mulanya PHP diciptakan dari ide Rasmus Lerdof yang membuat sebuah *script* perl. *Script* tersebut sebenarnya dimaksudkan untuk digunakan sebagai program untuk dirinya sendiri. Akan tetapi, kemudian dikembangkan lagi sehingga menjadi sebuah bahasa yang disebut "*Personal Home Page*". Inilah awal mula munculnya PHP sampai saat ini.

2.3. Basis Data

Basis data didefinisikan sebagai kumpulan data yang disatukan di dalam suatu organisasi. Basis data merupakan susunan / kumpulan data operasional lengkap dari suatu organisasi / perusahaan yang diorganisir/dikelola dan

disimpan secara terintegrasi dengan menggunakan metode tertentu, yaitu menggunakan komputer sehingga mampu menyediakan informasi yang optimal sesuai yang dibutuhkan pemakai.

Menurut Lukmanul Hakim (2009:10) pengertian Basis Data (*Database*) adalah: *"Kumpulan fiel-file yang mempunyai kaitan antara satu file dengan file lain sehingga membentuk satu bangunan data untuk menginformasikan suatu perusahaan instansi, dalam batasan tertentu".*

Dari pengertian tersebut dapat diambil kesimpulan bahwa Basis Data (*Database*) merupakan kumpulan dari data yang saling berhubungan satu dengan yang lainnya, tersimpan atau disimpan komputer dan digunakan perangkat lunak untuk memanipulasinya.

2.3.1. MYSQL

MySQL adalah sebuah perangkat lunak sistem manajemen basis data SQL (bahasa Inggris: *database management system*) atau DBMS yang multithread, multi-user, dengan sekitar 6 juta instalasi di seluruh dunia. MySQL AB membuat MySQL tersedia sebagai perangkat lunak gratis dibawah lisensi GNU *General Public License (GPL)*, tetapi mereka juga menjual dibawah lisensi komersial untuk kasus-kasus dimana penggunaannya tidak cocok dengan penggunaan GPL.

Tidak sama dengan proyek-proyek seperti *Apache*, dimana perangkat lunak dikembangkan oleh komunitas umum, dan hak cipta untuk kode sumber dimiliki oleh penulisnya masing-masing, MySQL dimiliki dan disponsori oleh sebuah perusahaan komersial Swedia MySQL AB, dimana memegang hak cipta hampir atas semua kode sumbernya. Kedua orang Swedia dan satu orang Finlandia yang mendirikan MySQL AB adalah David Axmark, Allan Larsson, dan Michael "Monty" Widenius.

MySQL merupakan implementasi dari sistem manajemen basis data relasional (RDBMS) yang didistribusikan secara gratis dibawah lisensi GPL

(General Public License). Setiap pengguna dapat secara bebas menggunakan MySQL, namun dengan batasan perangkat lunak tersebut tidak boleh dijadikan produk turunan yang bersifat komersial. *MySQL* sebenarnya merupakan turunan salah satu konsep utama dalam basisdata yang telah ada sebelumnya, SQL *(Structured Query Language)*. SQL adalah sebuah konsep pengoperasian basis data, terutama untuk pemilihan atau seleksi dan pemasukan data, yang memungkinkan pengoperasian data dikerjakan dengan mudah secara otomatis.

2.4. Text Mining

Text mining (penambangan teks) adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda beda (Feldman dan Sanger, 2007). *Text mining* merupakan teknik yang digunakan untuk menangani masalah klasifikasi, *clustering*, *information extraction dan information retrival* (Berry dan Kogan, 2010). Pada dasarnya proses kerja dari *Text mining* banyak mengapdopsi dari penelitian *Data Mining* namun yang menjadi perbedaan adalah pola yang digunakan oleh *text mining* diambil dari sekumpulan bahasa alami yang tidak terstruktur sedangkan dalam *Data Mining* pola yang diambil dari database yang terstruktur (Han dan Kamber, 2006). Tahap-tahap *Text mining* secara umum adalah *text preprocessing* dan *feature selection* (Feldman dan Sanger 2007, Berry dan Kogan 2010) . Dimana penjelasan dari tahap-tahap tersebut adalah sebagai berikut:

2.4.1. Text Preprocessing

Tahap *text preprocessing* adalah tahap awal dari *text mining*. Tahap ini mencakup semua rutinitas, dan proses untuk mempersiapkan data yang

akan digunakan pada operasi *knowledge discovery* sistem *text mining* (Feldman dan Sanger, 2007). Tindakan yang dilakukan pada tahap ini adalah toLowerCase, yaitu mengubah semua karakter huruf menjadi huruf kecil dan *Tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter delimiter seperti tanda titik (.), koma (,), spasi dan karakter angka yang ada pada kata tersebut.

2.4.2. Feature Selection

Tahap seleksi fitur (*feature selection*) bertujuan untuk mengurangi dimensi dari suatu kumpulan teks, atau dengan kata lain menghapus kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dokumen sehingga proses pengklasifikasian lebih efektif dan akurat (Feldman dan Sanger, 2007., Berry dan Kogan, 2010). Pada tahap ini tindakan yang dilakukan adalah menghilangkan stopword (*stopword removal*) dan *stemming* terhadap kata yang berimbuhan (Berry dan Kogan, 2010., Feldman dan Sanger, 2007). *Stopword* adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen (Dragut et al. 2009). Sebelum proses *stopword removal* dilakukan, harus dibuat daftar *stopword* (*stoplist*). Jika termasuk di dalam *stoplist* maka kata-kata tersebut akan dihapus dari deskripsi sehingga kata-kata yang tersisa di dalam deskripsi dianggap sebagai kata-kata yang mencirikan isi dari suatu dokumen atau *keywords*. Setelah melalui proses *stopword removal* tindakan selanjutnya adalah yaitu proses *stemming*. *Stemming* adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya (*stem*). Tujuan dari proses *stemming* adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata. Jika imbuhan tersebut tidak dihilangkan maka setiap satu kata dasar akan disimpan dengan berbagai macam bentuk yang

berbeda sesuai dengan imbuhan yang melekatinya sehingga hal tersebut akan menambah beban *database*. Hal ini sangat berbeda jika menghilangkan imbuhan-imbuhan yang melekat dari setiap kata dasar, maka satu kata dasar akan disimpan sekali walaupun mungkin kata dasar tersebut pada sumber data sudah berubah dari bentuk aslinya dan mendapatkan berbagai macam imbuhan.

2.5. Sentiment Analysis

Sentiment analysis atau *Opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *text mining* yang bertujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu. Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral. *Sentiment analysis* juga dapat menyatakan perasaan emosional sedih, gembira, atau marah. Berdasarkan *Sentiment Analysis*, Kita dapat mencari pendapat tentang produk-produk, merk atau orang-orang dan menentukan apakah mereka bersifat positif atau negatif.

2.6. Nai’ve Bayes Classifier

Nai’ve Bayes Classifier merupakan sebuah metode klasifikasi yang berakar pada teorema bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris *Thomas Bayes,* Yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Ciri utama dari *Nai’ve Bayes Classifier* ini adalah asumsi yang sangat kuat akan independensi dari masing masing kondisi / kejadian.

Menurut Olson Delen (2008) menjelaskan *Nai've bayes* untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari master tabel keputusan.

Nai've bayes classifier bekerja sangat baik dibanding dengan model classifier lainnya. Hal ini dibuktikan oleh Xhemali, Hinde Stone dalam jurnalnya "*Nai've Bayes vs. Decision Tree vs. Neural Networks in the Classification of Training Web Pages*" mengatakan bahwa *Nai've bayes classifier* memiliki tingkat akurasi yang lebih baik dibanding model classifier lainnya.

Keuntungan penggunaan metode *Nai've bayes classifier* adalah metode ini hanya membutuhkan jumlah data pelatihan (*Training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian.

Karena yang diasumsikan sebagai *variable independent*, maka hanya varian dari suatu variable dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

Secara garis besar model NBC (*Nai've Bayes Classifier*) adalah sebagai berikut:

$$p(C|F1,\dots,Fn) = \frac{p(C)\,p(F1,\dots,Fn|C)}{p(F1,\dots,Fn)}. \qquad (2.1)$$

Dengan
F adalah Dokumen
C adalah Jenis klasifikasi

Kita akan mengklasifikasikan suatu dokumen berdasarkan isi / kata-kata yang ada dalam dokumen tersebut. Sebagai contoh adalah apakah sebuah dokumen tersebut merupakan dokumen terkait bidang pendidikan / tidak.
Untuk itu, kita bayangkan bahwa sebuah dokumen – dokumen diambil dari suatu kelas dokumen (*Class of document*) yang dapat dimodelkan sebagai sebuah himpunan kata-kata, dimana probabilitas (*independent*) bahwa suatu kata ke-I dalam suatu dokumen terdapat dalam sebuah dokumen yang berasal dari class C. hal tersebut dapat digambarkan dengan:

$$p(w_i|C) \tag{2.2}$$

Dengan
W adalah kata
C adalah klasifikasi

(Atau untuk memudahkannya dapat kita asumsikan bahwa probabilitas suatu kata dalam suatu dokumen adalah independen terhadap ukuran suatu dokumen, dengan kata lain semua dokumen diasumsikan berukuran sama).

Selanjutnya probabilitas bahwa sebuah dokumen D, terhadap class C adalah:

$$p(D|C) = \prod_i p(w_i|C) \tag{2.3}$$

Dengan
D adalah dokumen
C adalah jenis klasifikasi

Pertanyaannya adalah “Berapa probabilitas suatu dokumen D merupakan milik suatu Class C?”, dengan kata lain adalah berapa nilai probabilitas

$$p(C|D) \tag{2.4}$$

Dengan

C adalah jenis klasifikasi

D adalah dokumen

Berdasarkan aksioma probabilitas:

$$p(D|C) = \frac{p(D \cap C)}{p(C)} \tag{2.5}$$

dan

$$p(C|D) = \frac{p(D \cap C)}{p(D)} \tag{2.6}$$

Dengan

C adalah jenis klasifikasi

D adalah dokumen

Untuk menjawab permasalahan sebelumnya diatas, maka kita asumsikan bahwa hanya terdapat dua kelas, yaitu kelas Spam (S), Bukan Spam (~S). dengan demikian model dapat digambarkan menjadi:

$$p(D|S) = \prod_i p(w_i|S) \tag{2.7}$$

dan

$$p(D|\neg S) = \prod_i p(w_i|\neg S) \tag{2.8}$$

Dengan

D adalah dokumen

S adalah klasifikasi Spam

-S adalah klasifikasi bukan spam

W adalah kata

Dari teorema bayes tersebut diatas, dapat kita tuliskan menjadi:

$$p(S|D) = \frac{p(S)}{p(D)} \prod_i p(w_i|S) \quad (2.9)$$

Dengan

S adalah klasifikasi spam

D adalah Dokumen

W adalah kata

Akhirnya, dokumen dapat diklasifikasikan sebagai berikut, dokumen tersebut merupakan Spam apabila:

$$ln \frac{p(S|D)}{p(\neg S|D)} > 0 \quad (2.10)$$

Dengan

S adalah klasifikasi Spam

-S adalah klasifikasi bukan Spam

D adalah Dokumen

Dan sebaliknya apabila < 0, maka dokumen tersebut bukan Spam